CLOUDERA

DSCI-273: Enterprise AI with Cloudera Machine Learning

Using Generative AI and Large Language Models (LLMs) with Enterprise Data

Course Overview

Course Type

Instructor-led training course

Level

Intermediate

Duration

4 days

Platform

CDP Public Cloud
CDP Private Cloud

Topics Covered

- CML Projects and Sessions
- Machine Learning Workflow
- Applications and AMPs
- Experiments and Models
- Data Visualization
- Large Language Models (LLMs)
- Prompt Engineering
- Retrieval Augmented Generation (RAG)
- Fine Tuning
- Merging LLM Models
- Metrics and Monitoring
- Performance and GPUs

About This Training

Generative AI (GenAI) and Large Language Models (LLMs) are extremely powerful new tools that are changing every industry. To fully take advantage of GenAI and LLMs, these new capabilities need to be combined with your existing enterprise data. This four-day course teaches how to use Cloudera Machine Learning to train, augment, and fine tune LLMs to create powerful enterprise AI solutions.

The course follows the Machine Learning Operations (MLOps) workflow to build enterprise machine learning applications. Participants learn how to explore and visualizing data, conduct experiments using MLFlow, use AMPs to accelerate solution development, deploy models as a REST API, and monitor model performance.

What Skills You Will Gain

Through lecture and hands-on exercises, you will learn how to:

- Utilize Cloudera SDX and other components of the Cloudera Data Platform to locate data for machine learning experiments
- Use an Applied ML Prototype (AMP)
- Manage machine learning experiments
- Connect to various data sources and explore data
- Select the right LLM model for a use case
- Configure a Prompt for an LLM
- Use Retrieval Augmented Generation (RAG)
- Fine Tune an LLM Model with Enterprise Data
- Deploy an ML model as a REST API
- Manage and monitor deployed ML models

Who Should Take This Course?

The course is designed for data scientists and machine learning engineers who need to understand how to utilize Cloudera Machine Learning and the Cloudera Data Platform to leverage the full power of their enterprise data, generative AI, and Large Language Models to deliver powerful business solutions.

Other Training That Might Interest You

- *Introducing Python*
- *Introducing Git*

DSCI-273: Enterprise AI with Cloudera Machine Learning

Training Outline (Page 2 of 3)

Introduction to CML

- Overview
- CML Versus CDSW
- ML Workspaces
- Workspace Roles
- Projects and Teams
- Settings
- Runtimes/Legacy Engines
- Lab: Introduction to CML

Introduction to AMPs and the Workbench

- Editors and IDE
- Git
- Embedded Web Applications
- AMPs
- Lab: Streamlit

Data Access and Lineage

- SDX Overview
- Data Catalog
- Authorization
- Lineage
- Lab: Data Access

Data Visualization in CML

- Data Visualization Overview
- CDP Data Visualization Concepts
- Using Data Visualization in CML
- Lab: Data Visualization

Experiments

- Experiments in CML
- Lab: Experiment Tracking

Introduction to LLMs

- History of LLMs
- How Transformers Work
- Different Types of LLMs
- Limitations of LLMs

LLM Model Selection

- How LLMs are Evaluated
- Model Selection by Use Case
- Hugging Face Model Hub
- Demo/Lab - Open LLM Leaderboard
- Demo/Lab - Can you run it? LLM Version

Prompt Engineering

- Components of a Prompt
- Shot Prompting
- Demo/Lab - Code Lama Playground 13B
- Demo/Lab - Mistral 7B Instruct

Text Summarization with Amazon Bedrock

- Amazon Bedrock Key Features
- Amazon Bedrock Use Cases
- Demo/Lab - Text Summarization with Amazon Bedrock

Retrieval Augmented Generation

- Retrieval Augmented Generation (RAG)
- RAG Use Cases
- Demo/Lab - LLM Chatbot Augmented with Enterprise Data
- Demo/Lab - Intelligent QA Chatbot with NiFi, Pinecode, and Llama2

DSCI-273: Enterprise AI with Cloudera Machine Learning

Training Outline (Page 3 of 3)

Fine Tuning

- Motivation for Fine Tuning
- Principles of Fine Tuning

Parameter Efficient Tuning

- Limitations of Fine Tuning
- Principles of Parameter Efficient Tuning

Fine Tuning a Foundation Model

- Quantization
- Low Rank Adaptation
- Demo/Lab - Fine Tuning a Foundation Model for Multiple Tasks (with QLoRA)

Merging LLM Models

- LLM Merging Core Principles
- Merging Benefits and Potential Applications

Deploying a Machine Learning Model as a REST API in CML

- Load the Serialized Model
- Define a Wrapper Function to Generate a Prediction
- Test the Function

Autoscaling, Performance, and GPU Settings

- Autoscaling Workloads
- Working with GPUs
- Lab: Autoscaling

Model Metrics and Monitoring

- Why Monitor Models?
- Common Models Metrics
- Models Monitoring with Evidently
- Continuous Model Monitoring
- Lab: Model Monitoring